Development Engineering 4 (2019) 100038

Contents lists available at ScienceDirect

Development Engineering

journal homepage: www.elsevier.com/locate/deveng

Assessing the causal impact of Chinese aid on vegetative land cover in Burundi and Rwanda under conditions of spatial imprecision

Robert Marty[a], Seth Goodman[b], Michael LeFew[b], Carrie Dolan[c], Ariel BenYishay[b,d], Daniel Runfola[b,e,*]

[a] The World Bank, 1818 H St NW, Washington, DC 20433, USA
[b] The Institute for Global Research, AidData, The College of William and Mary, 424 Scotland Street, Williamsburg, VA 23185, USA
[c] Department of Kinesiology and Health Sciences, The College of William and Mary, PO Box 8795, Williamsburg, VA 23187, USA
[d] Department of Economics, The College of William and Mary, PO Box 8795, Williamsburg, VA 23187, USA
[e] Department of Applied Science, The College of William and Mary, PO Box 8795, Williamsburg, VA 23187, USA

ARTICLE INFO

Keywords:
Aid
International development
Spatial imprecision
Geoparsing
Forest cover

ABSTRACT

There has been considerable debate regarding the efficacy of international aid in meeting the dual goals of human development and environmental sustainability. Many donors have sought to engage with this challenge by introducing environmental safeguard and monitoring initiatives; however, evidence on the success of these interventions is limited. Evaluating aid is a particular challenge in the case of donors that do not disclose information on the nature, geographic location, or extents of their interventions. In such cases, new methods that extract and geoparse data on the activities of opaque donors through the manual interpretation of thousands of news and other articles allow us to investigate the impacts of these activities. However, residual spatial uncertainty in these data remains a potential source of bias. In this article, we apply and discuss a Geographic Simulation and Extrapolation (GeoSIMEX) approach to mitigate the spatial imprecision inherent in geoparsed data. In conjunction with GeoSIMEX, we test and contrast multiple approaches to reducing the imprecision of aid, including high-assumption cases in which other covariates (i.e., nighttime lights) are leveraged to allocate aid. In our application, we find that methods which do not account for spatial imprecision find statistically significant relationships between Chinese aid and vegetation change; after accounting for spatial uncertainty, findings are similar for Rwanda and inconclusive for Burundi.

1. Introduction

Following recent calls to action from the United Nation's Intergovernmental Panel on Climate Change (McCarthy, 2001; Romero-Lankao et al., 2014) and multiple other organizations, international donors are being challenged to ensure that human development goals are not achieved in ways that may threaten environmental sustainability (Kareiva et al., 2008; McShane et al., 2011). Different donors have sought to engage with this challenge in different ways, such as the monitoring and safeguards implemented by the World Bank (Buchanan et al., 2016). By examining international aid as a causal driver of vegetative land cover, we argue it is possible to provide a unique window into the effectiveness of donors in mediating their impact on the environment (c.f. Rindfuss et al., 2004; Turner et al., 2007; Turner et al., 2003).

Past studies have illustrated the potential of using geoparsed data consisting of the latitude and longitude of aid interventions combined with satellite imagery to understand the impacts of aid on the environment (Buntaine et al., 2015; BenYishay et al., 2017; Runfola et al., 2017; Zhao et al., 2017a, 2017b; Buchanan et al., 2016). These studies, relying on a mixture of quasi-observational matching and difference in difference modeling strategies, have examined topics ranging from the protection of indigienous lands to infrastructure development, and related impacts on the environment. However, in every case, relatively precise geospatial data on the location of aid was provided or retrieved in collaboration with the donor organization.

Even in these cases, researchers frequently chose to remove - or otherwise permute - some observations from analyses due to a lack of fine detail geospatial data. This spatial imprecision inherent in such data presents a core limitation in the expansion of these studies beyond narrow contexts. In this piece, we provide an application of the Geographic Simulation and Extrapolation (GeoSIMEX) approach that

* Corresponding author. Data Science Program, The College of William and Mary, PO Box 8795 Williamsburg, VA 23185, USA.
E-mail address: danr@wm.edu (D. Runfola).

https://doi.org/10.1016/j.deveng.2018.11.001
Received 15 December 2017; Received in revised form 20 September 2018; Accepted 12 November 2018
Available online 23 November 2018
2352-7285/ © 2018 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

mitigates the potential bias from spatial imprecision. We do this in the context of a specific application, presenting an illustrative analysis of the impact of Chinese infrastructure aid on vegetative land cover in Rwanda and Burundi.

As an "opaque" donor, China does not formally report recipients of aid or development finance; project location information is not made publicly available, if it exists at all. Following this, here we leverage a recently-published, geoparsed dataset produced by the AidData research lab (AidData, 2017). This dataset was produced through a manual geoparsing procedure, in which place names were recorded and assigned latitude and longitude information when it is available. We focus on Rwanda and Burundi because both settings contain varying degrees of spatial imprecision in their geoparsed data.

Because of inherent limitations in the available documentation, it is common for the geographic information that is provided to be spatially imprecise - i.e., rather than knowing the exact location of an intervention, only the town or county in which it occurred is known. To mediate this challenge, we leverage GeoSIMEX in which spatial imprecision is treated as a special case of error-in-variables. Further, we contrast commonly employed approaches to allocating aid associated with imprecisely defined locations, including a comparison of the use of population density and nighttime lights to determine allocation. In our application, we find that methods which do not account for spatial imprecision find statistically significant relationships between Chinese aid[1] and vegetation change; after accounting for spatial uncertainty, findings are similar for Rwanda (low spatial imprecision in aid dataset) and inconclusive for Burundi (high spatial imprecision in aid dataset).

2. Theory

2.1. International development and impacts on the environment

There has been considerable debate in the literature regarding the efficacy of international aid in supporting human development - and the associated environmental costs (c.f. Kapur et al., 1997; McKibbin and Wilcoxen, 2002; Wade, 2003; Buntaine et al., 2015; Shandra, 2007). For indicators of economic growth, the cross-national literature ranges from pointing to aid having no association with growth (Rajan and Subramanian, 2008) to aid having a positive impact on growth in countries with policies conducive to growth (Dollar and Levin, 2005). Using nighttime lights as a proxy for economic growth, subnational analyses have found strong correlations between World Bank aid and nighttime lights, but no correlation between Chinese aid and nighttime lights (Dreher et al., 2015; Isaksson and Kotsadam, 2016). Results are similarly mixed on the impact of aid on environmental indicators. Recent work has identified causal relationships between aid contributed by non-governmental organizations and reductions in deforestation (Shandra, 2007), but analyses focusing on the World Bank find that some typologies of projects have contributed to deforestation (Shandra et al., 2011). Sub-national evaluations of World Bank aid using more precise geospatial data (as opposed to country-level aggregates) have come to different conclusions. Buchanan et al., (2016) finds that biodiverse areas closer to World Bank projects experienced marginally less forest loss than comparable areas far from Bank projects; Runfola et al., (2017) finds considerable spatial heterogeneity in this relationship. Further, approaches evaluating specific projects have found some projects intending to reduce forest loss have not impacted relevant outcome metrics sufficiently to achieve statistically significant results (Buntaine et al., 2015; BenYishay et al., 2017).

These differential outcomes speak to the challenges of examining land change processes that differ across geography - both human and physical - and highlight recent calls to the land change science community to focus on approaches which both (a) improve our ability to understand the cumulative impact of many small events, and (b) further bridge the gap between local and global scale processes (see Foley et al., 2005; Nagendra et al., 2004; Turner et al., 2007). Although examining the causal impact of international aid on environmental outcomes has been a central goal of many communities, there has been a limited collaboration between geographers and economists to identify strategies to leverage spatial data for causally-identified analyses (Corrado and Fingleton, 2012). This has resulted in methodological limitations which stem from distinctions between modeling efforts seeking to predict relationships commonly taught and accepted by the geographic community (i.e., spatial regression or classification trees), and efforts which seek to establish causal relationships similarly taught and accepted by the economics community (i.e., propensity score matching or difference-in -difference modeling). Recent efforts have sought to merge these disciplinary approaches (see Drukker et al., 2013; BenYishay et al., 2017; Buntaine et al., 2015; Buchanan et al., 2016), but many open questions remain as to how concerns of spatial imprecision, autocorrelative effects and geographic heterogeneity can be mitigated (Runfola and Napier, 2016; Corrado and Fingleton, 2012). This paper engages explicitly with the case of spatial imprecision inherent in many sources of geoparsed data, and provides one methodological approach (GeoSIMEX) to overcome challenges of imprecision.

2.2. Spatial imprecision

The lack of exact geographic information on where measurements are obtained presents a frequent barrier to research. This has become increasingly evident as more scholars integrate geographic data from multiple sources - for example, census, satellite, and GPS sources - to try and establish causal or predictive relationships (c.f., Bare et al., 2015; Buntaine et al., 2015; Gallo and Goodchild, 2012; Andam et al., 2008). Past literature has shown that uncertainty in the locations of where measurements are taken can produce biased estimates in empirical analyses (Perez-Heydrich et al., 2013; Rettie and McLoughlin, 1999). For example, Perez-Heydrich et al. show that regression coefficients will be biased when using raster data in conjunction with point data, where the true locations of the point data are only known to exist within some 5–10 km radius of the measured location (Perez-Heydrich et al., 2013). Moreover, when outcome data are spatially joined to imprecisely measured aid locations, the resulting errors can lead to attenuation bias (Grilliches and Hausman, 1986). One frequently cited "best practice" to overcome this challenge is to take average raster values within a buffer encompassing where the point could have fallen, instead of the single raster value associated with the point (Perez-Heydrich et al., 2013; Rettie and McLoughlin, 1999). Another practice to address spatial uncertainty is to aggregate to some higher spatial scale where there is no spatial uncertainty (Perez-Heydrich et al., 2013). Yet another method is to only use information for which exact (or, otherwise very precise) geographic information is known (Runfola and Napier, 2016). Relatively little guidance exists as to what the impact of these choices might be.

Scholars in the geographic community have also engaged in research seeking to integrate information on the precision of geographic data to improve the accuracy of modeling efforts, focusing on avoiding biases that might manifest following aspatial approaches (Aerts et al., 2003; Ogryczak and Sliwinski, 2009). The approach used in this paper - GeoSIMEX - expands on these methods, integrating an extrapolation step into traditional Monte Carlo approaches to avoid bias due to imprecision. Such bias can be simply understood using a hypothetical case in which all data is fully imprecise - i.e., interventions are only known to have occurred somewhere within the study area, but it is unknown where. In a Monte Carlo approach, the expected coefficient generated by re-fitting such a model thousands of times would be 0 - representative of random noise. As spatial imprecision increases, so too

[1] In this paper, we examine all sources and types of Chinese official financing. We use the term "Chinese aid" to refer to these financial flows for ease of exposition.

will the bias towards 0 in Monte Carlo-based procedures; this is frequently referred to as attenuation bias.

In this piece, we illustrate the ways in which GeoSIMEX leverages extrapolation to mitigate attenuation bias due to spatial imprecision, and highlight how the performance of GeoSIMEX is mediated by imputation assumptions related to the geographic location of international aid (i.e., allocating aid towards population centers). Throughout each section we provide readers with insights into the key decisions that must be made when using imprecise spatial data, the implications of these decisions, and some of the tradeoffs in using GeoSIMEX relative to existing solutions. We then illustrate these tradeoffs in a case study focusing on the allocation of Chinese Aid to both Burundi and Rwanda. Finally, the Appendix provides detailed notes on a simulation study illustrating the effectiveness of GeoSIMEX, and guidance on the parameterization of GeoSIMEX in different hypothetical study contexts.

3. Calculation

In this section, we detail two common "best practices" to mitigating spatial imprecision - (1) excluding spatially imprecise data, and (2) using assumptions about aid allocation - and discuss the challenges related to these procedures.

3.1. Excluding spatially imprecise data

One approach frequently applied by researchers when faced with a dataset that includes data of varying spatial precision is to remove data which does not meet a spatial precision threshold, relying only on geographic locations that are known with a specified level of certainty. Excluding spatially imprecise aid avoids making explicit assumptions about how imprecisely measured interventions are truly distributed across space. This approach comes with a number of limitations, and can result in erroneous parameter estimates.

Excluding spatially imprecise aid may bias estimates of aid impacts if imprecise aid is allocated according to different processes than spatially precise aid. Such a dynamic may often exist in practice. For example, aid dedicated to sectors such as health frequently contain a mix of precision in documentation due to the need for confidentiality in certain classes of projects. Another example is when two different donors are active in the same country, but one donor reports very specific geographic information and another little to no geographic information. This type of systematic bias can result in erroneous parameter estimates.

Fig. 1 uses a hypothetical example of aid allocated to Uganda to illustrates the consequences of excluding spatially imprecise aid. Two coordinating donors allocate aid to Uganda: donor A allocates funds to all districts in Western Uganda and donor B allocates funds to all districts in Northern Uganda. However, donor B's project documentation is poor - donor B only specifies that aid was allocated somewhere in Uganda, and so the researcher has no way of knowing where this aid was allocated.

Using this example, three plausible scenarios are examined where the true impact of donor projects vary. In each scenario, using the unit of observation as each district, the following model is estimated:

$$Change\ in\ Wealth_i = \beta_0 + \theta_i Aid_i \quad (1)$$

where *Change in Wealth* is the change in wealth between two arbitrary years across each district, *Aid* is a binary variable indicating whether a district received aid, i is a unique index for each district, β_0 is the intercept (set to 1 for this example) and θ_i is the impact of aid in district i (set from one to four in this example, following a uniform random distribution). In this hypothetical, districts which receive no aid have a change in wealth of one unit in the positive direction, and units which receive aid can have a change in wealth from between 1 and 5 depending on the simulated impact of aid.

In the first scenario, both donor A and B's projects were effective, resulting in increases in wealth across most intervention areas - the map in Fig. 1b shows the hypothetical θ_i used for each district i. In the second scenario, only donor A's projects effectively increases wealth (Fig. 1c), and in the third scenario only donor B's projects effectively increases wealth (Fig. 1d).

Two estimates of *Aid* are contrasted for each scenario. First, the "true" locations of aid are used in a simple linear estimation (following eq. (1)), where districts that received aid from either donor A or donor B are considered treated. The second estimate is representative of what a researcher might actually observe, in which only locations with spatially precise aid are used for estimation. The goal in each case is the same: to accurately model the average θ (i.e., estimate the average treatment effect) in the "true aid" case.

Fig. 1e–g illustrate the difference in coefficients estimates (θ) when the two cases (no spatial imprecision and removing spatially imprecise data) are compared. Because no error is introduced in equation (1), the mean estimates represented by "true" locations are representative of the true average θ across all units i. Differences between the "true" and spatially precise-aid only case represent the bias resultant from dropping out imprecise aid in each scenario.

When both donor A and B projects are effective (scenario 1; Fig. 1e), only using spatially precise aid results in a downward bias; further, the true impact is not captured within the 95% confidence interval. In contrast, when only donor A projects are effective (scenario 2; Fig. 1f), only using spatially precise aid results in an upward bias (noting the confidence interval still fails to include the true impact). When only donor B's projects are effective (scenario 3; Fig. 1g), only using spatially precise aid results in a strong downward bias (flipping the sign on θ).

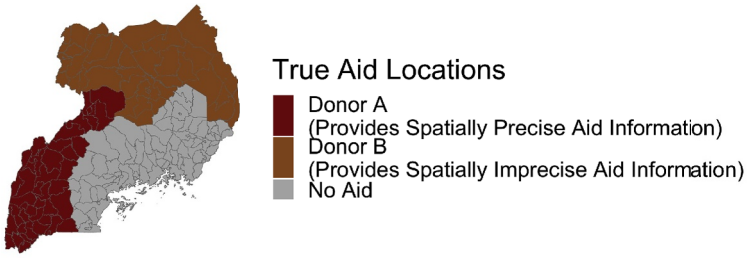
Because researchers are unaware as to the relative effectiveness of imprecise aid compared to precise aid, the direction of biases from excluding spatially imprecise data remains unknown in any research application. Spatially imprecise data may act as an omitted variable, as the true locations of spatially imprecise data may be correlated with both spatially precise data and the outcome of interest.

3.1.1. Assuming aid allocation process

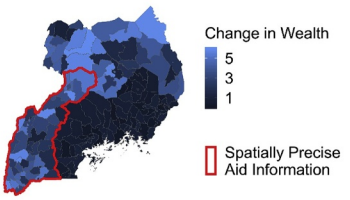
Another common approach to incorporating spatially imprecise data is to make assumptions regarding how data should be allocated to smaller spatial scales - a process referred to as both imputation (e.g., Jones et al., 2010; Sreenivas et al., 2014) and dasymetric mapping (e.g., (e.g., Holt et al., 2004; Bielecka, 2005; Eicher and Brewer, 2001) This section illustrates that the amount of bias that can occur from imputing aid to smaller spatial scales is a function of (1) the magnitude of spatial imprecision and (2) the accuracy of assumptions used for imputation.

To explore this, a hypothetical country with 60 subcounties, 30 counties, 10 districts and 2 regions is generated. Each subcounty is first randomly assigned a "true" probability of receiving aid (unknown to the researcher). Second, three hypothetical "imputed" probabilities of receiving aid (i.e., a probability that a researcher would generate) are generated: (1) A zero-correlation case, in which there is no correlation with the true probability; (2) a negative-correlation case, in which there is an inverse correlation with the true probability, and (3) a perfect correlation case, in which the researcher perfectly replicates the true aid probability. After these probabilities are assigned, 200 aid projects are randomly allocated to subcounties according to the true probability of receiving aid. In this hypothetical, each aid project is allocated $1 million USD towards a project designed to increase wealth. A $Total - Change - In - Wealth$ variable is then defined for each subcounty S as a one-to-one relationship according to equation (2):

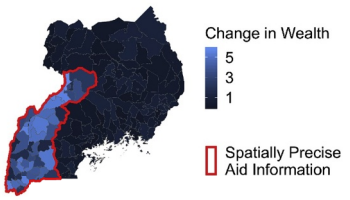
$$Total\ Change\ In\ Wealth_s = \theta Aid_s + \varepsilon \quad (2)$$



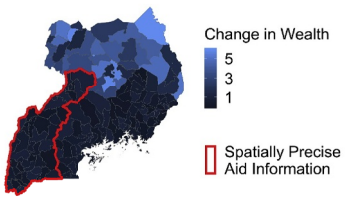
(a) Districts to which two hypothetical donors gave development aid.



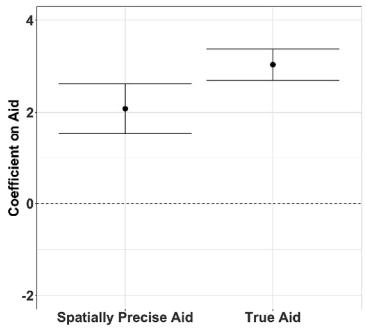
(b) Scenario 1: Hypothetical outcome if both donors aid results in an increase in wealth



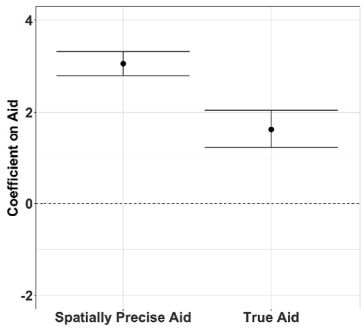
(c) Scenario 2: Hypothetical outcome if only donor A's aid results in an increase in wealth



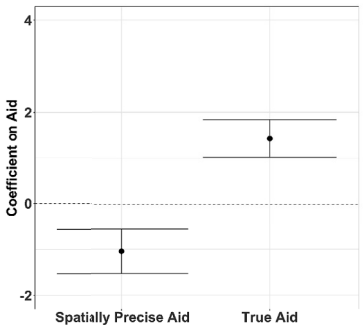
(d) Scenario 3: Hypothetical outcome if only donor B's aid results in an increase in wealth



(e) Scenario 1: Estimate of the effectiveness of aid under full-information ("True Aid") and limited information ("Spatially Precise Aid") conditions



(f) Scenario 2: Estimate of the effectiveness of aid under full-information ("True Aid") and limited information ("Spatially Precise Aid") conditions



(g) Scenario 3: Estimate of the effectiveness of aid under full-information ("True Aid") and limited information ("Spatially Precise Aid") conditions

Fig. 1. Comparison of different cases in which imprecise spatial information is excluded from the analysis.

where ε is normally distributed, and $\theta = 1$.

Spatial imprecision is then imposed on a random subset of aid projects by masking the true subcounty where the project was allocated to replicate the data a researcher might have access to. After masking the true location, we re-estimate equation (2), calculating Aid_s following the three imputation scenarios outlined above (perfect, zero, and inverse correlation). Here, the amount of aid per subcounty involves imputing aid amounts from high-level administrative areas (i.e., districts, regions or the country) down to subcounties for spatially imprecise aid. For example, if a $1 million project is allocated to a district with four sub-counties, the $1 million is allocated proportionally based on the assumed probability within the subcounties. Consequently, as spatial imprecision grows, the aid variable will becoming increasing correlated with the assumed probability, up to an extreme case of perfect imprecision in which all aid allocation is determined by researcher-defined assumptions; only the total quantity of aid would be provided by the source data in this perfect imprecision scenario.

Fig. 2 illustrates how bias in regression coefficients changes as spatial imprecision increases under the three assumption scenarios. In each of the three figures, the x-axis represents spatial imprecision (with larger values indicating increasing imprecision, estimated following equation (3)), and the y-axis represents the estimated θ value (in which the true $\theta = 1$). The x-axis (λ) captures the degree of spatial imprecision in a dataset by calculating:

$$\lambda = \frac{\sum_i^P Area\ of\ Coverage}{\sum_i^P Total\ Possible\ Area\ of\ Coverage} \tag{3}$$

where i is an individual aid project out of P total aid projects, $Area\ of\ Coverage_i$ is project i's known area of coverage defined by the available documentation - i.e., the geographic area within which the project is known to have been implemented (i.e., the town in which a clinic was built), and $Total\ Possible\ Area\ of\ Coverage_i$ is the area of coverage of project i under complete spatial imprecision. Conceptually, this

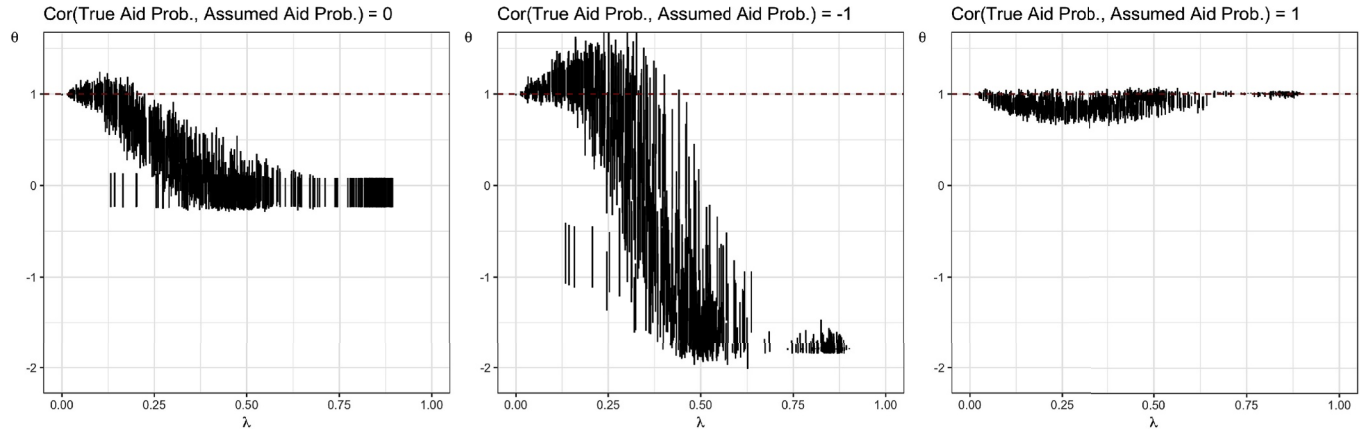


Fig. 2. Consequences of Spatial Imprecision. Each panel shows the relationship between a simulated true impact of 1 and estimated impacts, with vertical black bars representing naive (OLS) estimates of that impact (95 percent confidence intervals). The y-axis represents the estimated impact, and the x axis is the degree of spatial imprecision in the data. Panel 1 represents the case in which a researchers estimates of aid allocation are random (i.e., the "null" case), the second panel represents the case where estimates of aid allocation are incorrect (R = −1), and the third panel indicates the case where a researchers estimates of aid allocation are correct (R = 1).

represents the maximum possible geographic area across which aid might be distributed if you have no additional knowledge other than that it was distributed to a location within your study region. For example, if you are studying a particular country's aid allocation, it may be known that the federal government received aid, but no specific district within the country to which that aid was given is known. In this case, both the numerator and denominator would be the geographic area of the country. λ ranges between 0 and 1, in which 0 indicates perfect precision (i.e., exact knowledge of locations is known), and 1 indicates perfect imprecision (all that is known is that each aid intervention occurred somewhere within the entire study area).

100 simulations are shown on each plot in Fig. 2, and 95% confidence intervals of the θ estimate are provided. Panel 1 illustrates the increasing bias in aid under a random aid allocation scenario - as the precision of known aid decreases, by design the θ estimates approach 0. This panel illustrates the importance of both the imputation procedure chosen, as well as the spatial precision of the source data - under strong spatial precision cases, even random allocation strategies can still result in valid coefficient estimates. Panel 2 illustrates the variation in coefficient estimates which can occur if a researcher makes very bad assumptions about where aid is geographically allocated, represented by a correlation of −1. As this figure shows, strong source data can mitigate the impact of poor assumptions, but only to a λ of approximately 0.25. Panel 3 illustrates that, even in cases of extreme imprecision, distributional assumptions will not bias regression coefficients when assumptions are correct.

4. Data and methods

4.1. Data

As an illustrative case study incorporating spatial uncertainty, we examine the causal impact of Chinese infrastructure aid distributed from 2007 to 2011 on vegetation in Burundi and Rwanda. This study leverages a dataset on the location of Chinese foreign aid at varying levels of precision (i.e., the exact location of each project is not always known; AidData, 2017). Table 1 summarizes the number of projects and spatial imprecision of the Chinese data for Burundi and Rwanda. The Chinese aid data for Burundi has a moderate amount of spatial imprecision (λ = 0.37); over two-thirds of the project locations are measured at a level coarser than the unit of analysis (ADM2) for the study. Contrasting to this, the Chinese aid data for Rwanda is measured with mild spatial imprecision (λ = 0.15); 20 out of 27 projects are measured at the level of the unit of analysis.

Table 1
Summary of data for case studies.

Dataset Characteristic	Burundi	Rwanda
Number of Projects	9	13
Number of Project Locations	17	27
Spatial Precision: ADM 2 Level	5	20
Spatial Precision: ADM 1 Level	6	4
Spatial Precision: Country-Level	6	3
Spatial Imprecision (λ)	0.37	0.15

Vegetation is measured using a satellite-derived metric of vegetative density, the Normalized Difference Vegetation Index (NDVI), from NASA's Long Term Data Record (LTDR; NASA, 2017). While the goal of this case study is illustrative - i.e., we seek to contrast GeoSIMEX to other methodological approaches, we provide an example of how differences in findings may occur even in models that have the goal of causal attribution. Thus, a difference-in-difference modeling strategy is followed, in which average NDVI before aid was allocated (pre-2007) is contrasted to the average NDVI after aid was allocated (post-2011). The model is described in equation (4):

$$Y_i = \beta_0 + \theta A_i + \sum_{k=1}^{N} \beta_k X_i + \varepsilon_i \quad (4)$$

where Y_i is the difference in the average forest loss post-2011 (2012 and 2013) and pre-2007 (2001–2006) in a country's second administrative division (ADM2); θ represents the estimated impact of aid, X_i is a vector of controls which includes changes (i.e., post minus pre periods) in average amounts of nighttime lights and maximum, minimum, and average levels of air temperature and precipitation, k an index for each control covariate, A_i the aid allocated to each ADM2 and ε_i are the error terms.

After estimating OLS models, we test whether spatial autocorrelation exists in the residuals by calculating Moran's I. As an additional point of comparison with the existing literature on bias due to spatial autocorrelation (c.f. Anselin, 2003; Anselin, 2002; Anselin and Cho, 2002; Anselin, 2013; LeSage and Pace, 2009), we estimate spatial error models as described in equation (5)

$$\begin{aligned} Y_i &= \beta_0 + \theta A_i + \textstyle\sum_{k=1}^{N} \beta_k X_i + \varepsilon_i \\ \varepsilon_i &= \tau W \varepsilon_i + \xi_i \end{aligned} \quad (5)$$

where ε_i are the error terms spatially weighted using weights matrix W, τ is the spatial error coefficient and ξ_i are uncorrelated error terms. Each dataset is processed and aggregated according to average values within each ADM2.

A_i is calculated following four estimation strategies to examine the robustness of findings when different imputation approaches are followed. These approaches include allocating aid proportionally according to: (a) baseline (2005) values of population (CIESIN, 2000), (b) average nighttime lights (NOAA, 2017), (c) spatial area (GADM), and (d) the amount (in USD) of spatially precise aid allocated within each ADM2 (AidData, 2017). Four naive and GeoSIMEX models are estimated, one using each imputation approach, for a total of eight models.

4.2. GeoSIMEX method

This section describes the methodology employed to account for spatial imprecision in this paper —Geographic SIMEX (GeoSIMEX). We provide illustrative simulation results in Appendix A, as well as a discussion of the key ways in which our GeoSIMEX application in this piece diverges from previous applications. GeoSIMEX is based on the simulation and extrapolation method (SIMEX), which provides one solution to address measurement error in covariates (Wang et al., 1998; Li and Lin, 2003; Cook and Stefanski, 1994). It relies on estimating a trend between increasing spatial imprecision and bias, i.e., the trend observed in Fig. 2.

GeoSIMEX first involves estimating a naive model (which can be of variable functional forms; for illustration we use ordinary least squares regression). To estimate the naive model, aid measured with spatial imprecision must be imputed to the administrative level of the unit of observations (i.e., subcounties). Using the four imputation cases described above to estimate subcounty-level values of aid, a naive OLS model is estimated. In addition, the spatial imprecision (λ) value of the dataset is recorded. Additional error is simulated to re-create the dataset at coarser (larger) values of λ, the imputation process and model

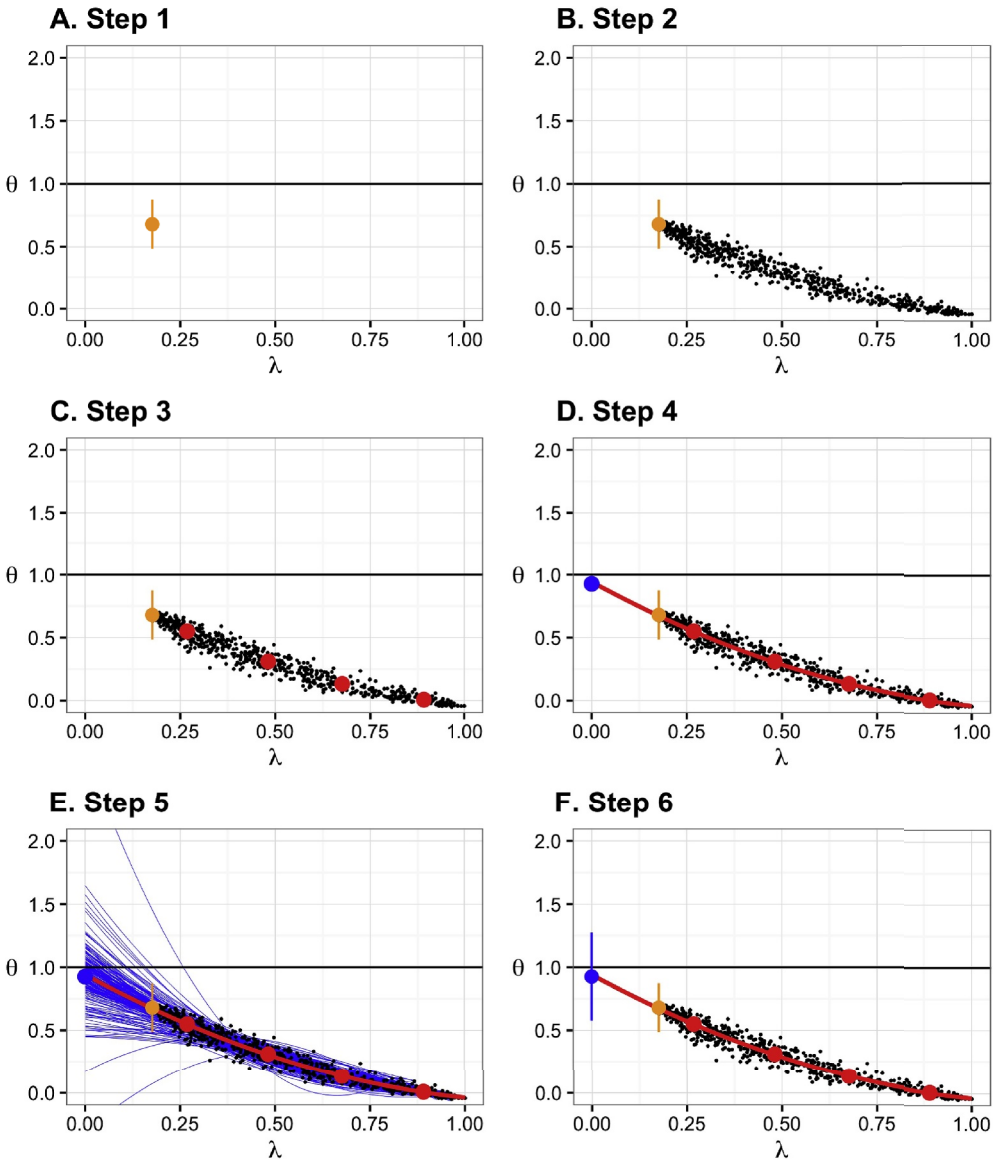


Fig. 3. Steps of GeoSIMEX. The Y-axis on each figure represents the treatment impact (θ) of a hypothetical case study. The X-axis on each figure represents the degree of spatial imprecision (λ) for each of the original (orange) and simulated (black) datasets.

fitting repeated, and parameter estimates and λ recorded for each imputation case. This information is then used to back-extrapolate for each imputation case to the point of zero imprecision, $\lambda = 0$.

The specific steps of GeoSIMEX are summarized in Fig. 3. Fig. 3A shows the estimation of the original model, based on the imputation strategy a researcher selected for the estimation of A_i (i.e., allocating aid to areas with higher populations). The point represents the coefficient estimate for θ, and vertical line indicates the 95% confidence interval. Each additional dot in Fig. 3B represents a model in which aid with a known level of spatial precision is intentionally decreased (i.e., additional imprecision is added), and the imputation and model fit procedure is repeated. Fig. 3C shows a binning procedure, in which an arbitrary number of bins are created from which parameter estimates can be sampled (more detail on the selection of these bins can be found in the below simulation analysis). Fig. 3D shows a back-extrapolation procedure, in which a best fit line is fit through the average values identified in each bin and the θ estimate is calculated for $\lambda = 0$ (i.e., the no spatial imprecision case). Fig. 3E shows a bootstrapped standard error procedure, in which a number of parameter estimates are taken from each bin and the extrapolation procedure repeated, resulting in a range of possible solutions. Finally, Fig. 3F shows the 95% confidence interval calculated based on the iterative extrapolation procedure in step 5 (Fig. 3E).

A number of parameters must be selected for the GeoSIMEX estimation procedure - we use a simulation procedure described in Appendix A to identify those that are most accurate in our use case. Key of these parameters is the number of bins used in the back-extrapolation procedure (Fig. 3C). As further detailed in Appendix A, for users seeking to use GeoSIMEX our simulation results suggest that the number of bins that is most appropriate is highly dependent on the spatial accuracy of the source data. For finer-resolution data, larger bin values (between 5 and 10) are more appropriate (see Table 4 for more specific guidance). For coarse-resolution data, smaller bin values (between 3 and 4) tend to provide more accurate results.

5. Results: case study

Tables 2 and 3 report results from the Burundi and Rwanda case studies respectively. Due to an indication of spatial autocorrelation in the case of Burundi, a Spatial Error Modeling approach was pursued follow equation (5). The Burundi case study illustrates how spatial imprecision can weaken researchers' ability to find robust, significant relations in data. The Burundi data has a moderate degree of spatial imprecision ($\lambda = 0.37$). In Burundi, naive models (1–4) show mixed results as to the impact of Chinese aid. When imprecise aid is allocated according to population, the naive model shows Chinese aid having a significant impact on reducing forest loss ($p < 0.01$); however, the other naive models show no significant association between Chinese aid and forest loss.

Contrasting to the more traditional modeling approaches, in Burundi none of the GeoSIMEX models show Chinese aid having a significant impact on forest loss. This is representative of an increase in the standard errors attributable to the incorporation of known spatial uncertainty, capturing the low spatial precision of this data.

The Rwanda case study illustrates how naive and GeoSIMEX models will be more similar under lower levels of spatial imprecision, as Rwanda has a relatively smaller amount of spatial imprecision ($\lambda = 0.15$) as contrasted to Burundi. Naive models (1–4) show that Chinese aid reduced forest loss across all models. When imprecise aid is allocated according to population, nighttime lights, or low precision aid, the coefficient on aid is significant at the 5% level; when imputing aid based on area the coefficient is significant at the 10% level.

Across different aid allocation assumptions, naive models show a relatively similar impact of Chinese aid; coefficients range from −0.008 to −0.012. When accounting for spatial imprecision through GeoSIMEX, the coefficients remain similar to those in the naive models. Additionally, when assuming aid is allocated according to population, nighttime lights or low precision aid, the coefficient on aid remains significant at the 5% level; when imputing aid based on area, the aid coefficient is no longer significant.

It is also useful to examine how standard errors change between naive and GeoSIMEX models. In the Rwanda case study, standard errors only mildly change due to low amounts of spatial imprecision (though they tend to increase). However, in the Burundi case study, standard errors change to different extents depending on assumptions regarding the allocation of spatially imprecise aid. When aid is allocated according to area, population, or low precision code aid, standard errors increase by 0.447, 0.114, and 0.401 respectively from naive to GeoSIMEX models. However, when aid is allocated according to nighttime lights, the standard error increases by 0.896. Simulation results suggests that less accurate assumptions when imputing A_i will yield larger standard errors; consequently, this suggests that nighttime lights may be an inaccurate assumption about the aid allocation process. However, dynamics beyond accuracy of assumptions influence changes in standard errors from naive to GeoSIMEX models, so any such interpretations must be conducted with caution.

Table 2
Case study results for Burundi under different assumptions of aid allocation. 'NTL' refers to baseline nighttime lights. 'PC' refers to precision code. OLS refers for ordinary least squares models and SEM refers to spatial error models. Control variables were included in all models; N = 109. ∗p < 0.1; ∗∗p < 0.05; ∗∗∗p < 0.01.

	Dependent variable: Percent Forest Loss											
	OLS				spatial error							
	Naive [OLS]				Naive [SEM]				geoSIMEX [SEM]			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Aid (Area)	0.03971				0.07826				−0.10045			
	(0.07511)				(0.07260)				(0.51943)			
Aid (Population)		−0.17876∗∗∗				−0.14301∗∗				−0.08245		
		(0.05779)				(0.05992)				(0.17440)		
Aid (NTL)			0.07193				0.09612				0.08979	
			(0.08123)				(0.07441)				(0.97009)	
Aid (Low PC Aid)				0.01351				0.02846				−0.24064
				(0.05208)				(0.04511)				(0.44589)

Table 3
Case study results for Rwanda under different assumptions of aid allocation. 'NTL' refers to baseline nighttime lights. 'PC' refers to precision code. OLS refers for ordinary least squares models. Control variables were included in all models; N = 30. ∗p < 0.1; ∗∗p < 0.05; ∗∗∗p < 0.01.

	Dependent variable: Percent Forest Loss							
	Naive [OLS]				geoSIMEX [OLS]			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Aid (Area)	−0.00942∗				−0.00836			
	(0.00504)				(0.00528)			
Aid (Population)		−0.01242∗∗				−0.01194∗∗		
		(0.00447)				(0.00456)		
Aid (NTL)			−0.01126∗∗				−0.01130∗∗	
			(0.00484)				(0.00514)	
Aid (Low PC Aid)				−0.00859∗∗				−0.00839∗∗
				(0.00384)				(0.00397)

6. Discussion and conclusion

Geographically-referenced aid information often is measured with spatial imprecision. When faced with spatial imprecision, researchers often exclude spatially imprecise data or make assumptions about how spatially imprecise aid was allocated. However, excluding imprecise data or making incorrect assumptions can lead to biased estimated coefficients, potentially resulting in misleading policy conclusions. This paper describes a method for accounting for spatial imprecision, GeoSIMEX, and illustrates how researchers can use it in conjunction with imputation assumptions for aid allocation processes.

The method has three notable advantages. First, the method can help correct biases in regression coefficients caused by spatial imprecision that Monte Carlo simulation alone can not. Second, the method increases uncertainty in regression coefficients as spatial imprecision grows, which allows results to incorporate the inherent quality of the underlying data (i.e., reflecting "known unknowns" of the data). Third, while the accuracy of assumptions about aid allocation processes will remain unknown to researchers, GeoSIMEX can be used in conjunction with imputation assumptions to enable researchers to trade off assumptions for higher degrees of spatial precision.

The presented case study illustrated how different assumptions about aid allocation can lead to different results, and highlights the importance of incorporating spatial imprecision in analyses. The Burundi case had moderate amounts of spatial imprecision, resulting in no GeoSIMEX models showing significant relationships between aid and deforestation, contrasting to a finding of statistical significance in one of the four test cases under a naive modeling case. In the case of Rwanda, which had more precise spatial data, the GeoSIMEX estimation strategy identified a significant relationship between Aid and Forest Loss in three of the four test cases; this contrasts only slightly to the naive modeling approach where all four of the test cases identified a significant relation. These findings are in-line with what one might intuitively expect: as the spatial data available becomes less precise, so too does our ability to make significant statements about underlying patterns.

Based on the findings of our case study and simulations, we suggest researchers facing similar challenges of spatial imprecision leverage methods such as GeoSIMEX over other existing tools (i.e., Monte Carlo analyses) to mitigate the chance of attenuation biases leading to erroneous results. However, the GeoSIMEX procedure requires researchers to make explicit a number of assumptions. Here, we provide some guidance for future researchers in how parameterizations and assumptions might be made in other cases, using our own case study as an example. We also note a number of potential avenues for future research:

- Parameterization of the number of bins used for back-extrapolation for GeoSimex. We calculate the spatial imprecision in our source data (λ), and use the guidance provided from simulations summarized in Table 5 to select an appropriate number of bins.
- Assumptions of how aid is allocated within units. In our case study results, we find that GeoSIMEX provides more stringent estimates of the impact of international aid irrespective of aid allocation decisions (i.e., significance is less likely to be found contrasted to naive models irrespective of the allocation assumption made). Simulation suggests that "null case" estimates (i.e., when allocation is conducted purely based on spatial area) provide a strong option, and due to the lack of additional assumptions is our preferred approach.
- There are a variety of options for the process of back-extrapolation to the $\lambda = 0$ point, as seen in Fig. 3D. Based on simulation results, we select a quadratic fit; future research could provide better guidance to researchers by contrasting more strategies.
- GeoSIMEX provides unbiased estimates in cases of minor to moderate spatial uncertainty (i.e., $\lambda \leq 0.5$), but has weaker results as the total spatial uncertainty increases beyond $\lambda > 0.5$ (see Fig. 5). This can be mediated by making strong assumptions regarding aid allocation, but is an avenue for future research (see Fig. 7 for simulation results examining this topic).

While we have explored many of these topics through simulation to identify the contexts in which GeoSIMEX assumptions and parameterizations are successful, we also recognize that the range of cases different researchers might engage with could extend beyond the scope of our simulations. Future research exploring more tailored simulations for a wider variety of contexts could help provide more explicit direction to researchers.

GeoSIMEX provides a step forward in allowing researchers to incorporate spatial imprecision into analysis. However, further work remains to improve GeoSIMEX. In particular, while GeoSIMEX reduces the risk of committing a type I error—wrongly concluding significance—it heightens the risk of committing a type II error—wrongly concluding insignificance. In this piece, we begin examining the potential of using ancillary information such as population and nighttime

lights to — in effect — tradeoff assumptions about where aid is allocated (i.e., to more populous regions) in exchange for a reduction in spatial imprecision. By exploring the potential for such tradeoffs, it is possible effective routes forward that can incorporate both spatial uncertainty and mitigate the risk of type II error may be found. Such approaches would have broad impact outside of research examining international aid, and have direct application to a wide range of practitioners following spatial imputation procedures to make fine-scale estimates based on spatially imprecise data.

Conflicts of interest

The authors declare no conflicts of interest.

Acknowledgements

This work was made possible by the support of the Global Environment Facility, the MacArthur Foundation, the Cloudera Foundation and the College of William and Mary.

This work was performed in part using computational facilities at the College of William and Mary which were provided with the assistance of the National Science Foundation, the Virginia Port Authority, Virginia's Commonwealth Technology Research Fund and the Office of Naval Research.

A. Simulation analysis of GeoSIMEX

A.1. Simulation methods

While previous research has begun to examine the accuracy of GeoSIMEX under "null assumption" cases (i.e., equally spreading aid across all subunits; Runfola et al., 2016), no simulation study has examined the efficacy of GeoSIMEX when different assumptions regarding the distribution of aid are made (e.g., distributing aid according to population). We employ a Monte Carlo simulation to examine the accuracy of GeoSIMEX at different levels of spatial imprecision and using a range of accuracies for the spatial imputation step (i.e., some simulations reflect a scenario where a researcher chooses an accurate imputation assumption, and others where they do not). Each simulation follows six steps.

First, one of three hypothetical countries with different administrative hierarchies is generated: (1) a country with 60 subcounties, 30 counties, 10 districts and 2 regions, (2) a country with 120 subcounties, 40 counties, 20 districts and 5 regions, or (3) a country with 120 subcounties, 60 counties, 30 districts and 10 regions. Each subcounty is randomly assigned a (known) probability of receiving aid. Additionally, to simulate the assumptions a researcher might make in the imputation of spatially imprecise aid, an assumed probability is generated which is correlated with the true probability by a random amount, ranging from -1 to 1.

Second, 50 to 250 aid projects are randomly allocated to subcounties, according to the assigned (known) probability of a subcounty receiving aid. **Third**, a simulated measurement of wealth is generated according to the following equation:

$$Wealth = \theta Aid + \varepsilon \tag{A.1}$$

where θ equals one, indicating there is a one-to-one relation between aid and wealth.

Fourth, each aid project is assigned a code indicating the spatial precision that a researcher might see in practice. Codes range from indicating that the project fell within a sub-county (no spatial imprecision) to falling somewhere within a country, district, region, or the country. Each code is assigned randomly, and the overall precision of the simulated dataset a researcher might see is quantified following equation (3) (λ).

Fifth, equation (A.1) is estimated using a linear model. Here, the expected value of aid is used—imputing aid based on the assumed probability of receiving aid for projects assigned a coarse precision. **Sixth**, equation (A.1) is estimated using two modeling approaches that explicitly incorporate spatial imprecision: GeoSIMEX and a Monte Carlo model averaging approach. In the model averaging approach, 500 regression models are estimated where spatially imprecise aid is imputed independently each iteration. At each iteration, each unit is assigned a random probability of receiving aid drawn from a 0–1 uniform distribution. Aid is imputed to units using this random probability. Coefficients and standard errors are averaged following Burnham and Anderson (2002).

This process is repeated approximately 5 million times to account for the large range of parameters being tested (see Table 4). Simulation results are examined in three steps. First, we identify the parameters within GeoSIMEX that optimize its performance according to two criteria: (1) the ability of GeoSIMEX to capture the true coefficient and (2) the ability of GeoSIMEX to capture the true coefficient with statistical significance at the 95% level.

Second, using only simulations which used the optimal parameters selected in step one, we test whether simulations corroborate the theory of GeoSIMEX. Specifically, we explore the intersection between imputation, imprecision and bias. Third, we compare the performance of GeoSIMEX to a naive model and the model averaging approach under both different levels of spatial uncertainty and accuracy of imputation assumptions.

A.2. Optimizing GeoSIMEX parameters

This section examines how the performance of GeoSIMEX varies with different parameters. Four parameters are tested within GeoSIMEX: (1) the number of iterations within GeoSIMEX where additional imprecision is simulated (100, 250 or 500), (2) the number of bins that are used to perform extrapolations, (3) the number of values taken from each bin to perform extrapolations, and (4) whether the values from each bin are averaged before taking an extrapolation. Two metrics of performance are examined: first, whether the 95% confidence interval on the aid variable captures the true coefficient and second, whether the 95% confidence interval captures both the true coefficient and statistical significance.

Linear regression models are estimated to understand how parameters are correlated with measures of performance. To explore how the optimal parameters for GeoSIMEX fluctuate based on the magnitude of spatial imprecision in a dataset, we divide the sample into three groups based on the level of spatial imprecision: $0 < \lambda < \frac{1}{3}$, $\frac{1}{3} < \lambda < \frac{2}{3}$, and $\frac{2}{3} < \lambda < 1$.

A.2.1. GeoSIMEX parameter optimization

Table 4 shows regression results for the parameter optimization. Each column in this table represents one of three cases. The first case (as seen in the first column of data) is representative of a regression in which a binary value (0 or 1) is the outcome (Y) value, and the parameters of each simulation iteration are regressed to establish which simulations resulted in outcome Y. In this first column, Y is equal to 1 in all simulation iterations in which, at the 95% confidence interval, the true estimate of treatment impact ($\theta = 0$) is observed. All other columns of data follow a similar approach; in the second column, Y is equal to 1 in all iterations in which the true estimate of the treatment impact ($\theta = 1$) is observed; in the third column $Y = 1$ when the true estimate of the treatment impact ($\theta = 1$) is observed and found to be statistically significant. Three groupings of these models are estimated, one for each level of spatial imprecision.

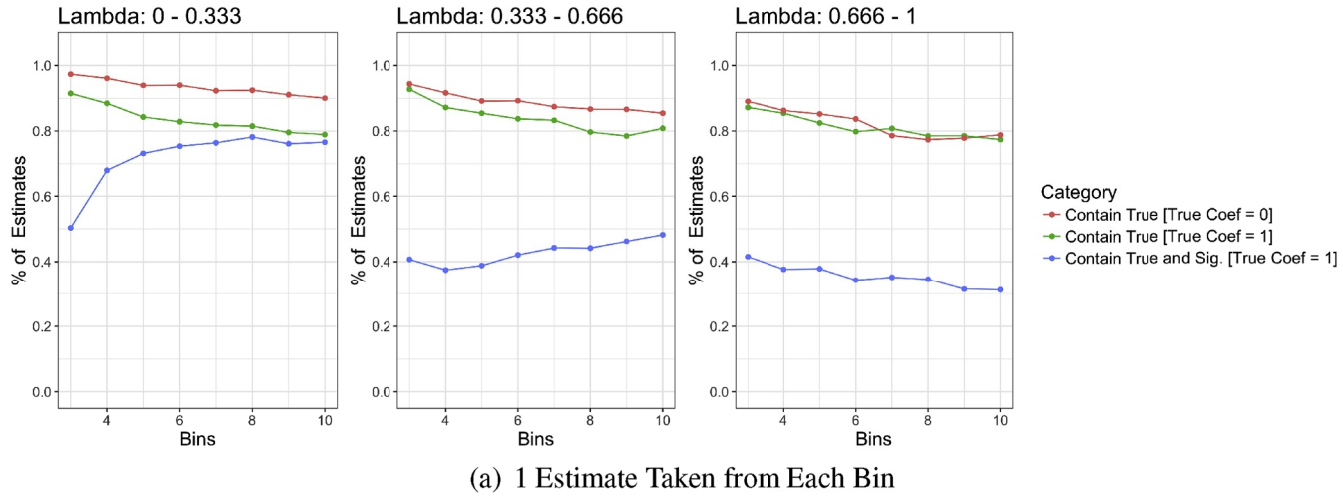
A few general trends emerge from these simulations. Across most models, a greater number of simulations within GeoSIMEX resulted in more accurate parameter estimates. Using a quadratic extrapolation versus a linear extrapolation also generally improves the performance of GeoSIMEX. This result suggests that users should first try using a quadratic extrapolation; however, users should always examine the trend between imprecision and regression coefficients to determine whether other methods of extrapolation would better fit the data.

The regression results show a notable tradeoff between the ability of GeoSIMEX to capture only the true coefficient or the true coefficient and significance in regards to the number of bins used and the number of values taken from within each bin. Greater number of bins used in the GeoSIMEX procedure tends to increase the ability of GeoSIMEX to capture the true coefficient; however, additional bins worsens the ability of GeoSIMEX to capture significance. Using greater number of values within each bin shows a similar trend of being beneficial to capture the true coefficient and significance but worsens the ability of GeoSIMEX to only capture the true coefficient. Within-bin averaging improves the ability of GeoSIMEX to capture the true coefficient in low to medium levels of spatial imprecision and worsens the ability of GeoSIMEX to capture significance under medium levels of spatial imprecision.

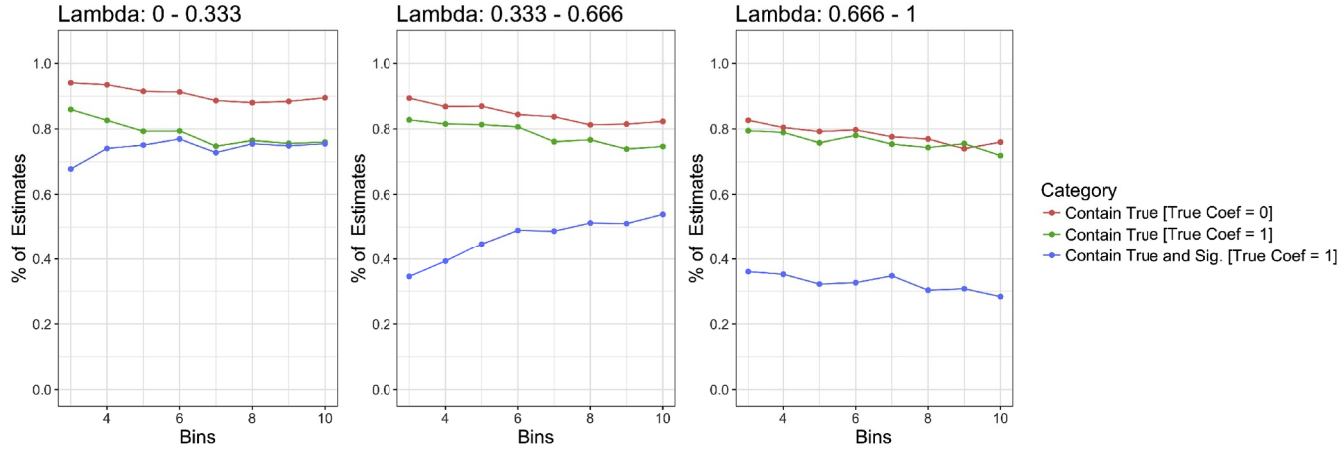
Table 4
Performance of GeoSIMEX Based on Parameters.

	$0 < \lambda < 0.33$			$0.33 < \lambda < 0.66$			$0.66 < \lambda < 1$		
	True Coef = 0	True Coef = 1		True Coef = 0	True Coef = 1		True Coef = 0	True Coef = 1	
	Contain True Coef	Contain True Coef	Contain True Coef &Sig.	Contain True Coef	Contain True Coef	Contain True Coef &Sig.	Contain True Coef	Contain True Coef	Contain True Coef & Sig.
250 Simulations	0.013***	0.014***	0.010***	0.022***	0.028***	−0.0001	0.031***	0.017***	−0.005
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
500 Simulations	0.014***	0.022***	0.017***	0.030***	0.035***	−0.006*	0.038***	0.026***	−0.003
	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
Sim. Precise Data	−0.072***	−0.191***	−0.253***	−0.073***	−0.084***	−0.167***	−0.038***	0.025***	0.012***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)
Quadratic Extra-p.	0.135***	0.281***	0.174***	0.275***	0.344***	−0.205***	0.428***	0.424***	−0.439***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)
Bins: 4	−0.014***	−0.032***	0.057***	−0.020***	−0.029***	0.032***	−0.031***	−0.034***	0.015***
	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Bins: 5	−0.029***	−0.059***	0.073***	−0.044***	−0.053***	0.069***	−0.042***	−0.061***	0.008
	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Bins: 6	−0.039***	−0.084***	0.067***	−0.057***	−0.078***	0.087***	−0.056***	−0.073***	0.006
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Bins: 7	−0.045***	−0.100***	0.058***	−0.062***	−0.092***	0.104***	−0.073***	−0.082***	0.012**
	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Bins: 8	−0.053***	−0.107***	0.060***	−0.075***	−0.109***	0.116***	−0.086***	−0.090***	0.008
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Bins: 9	−0.067***	−0.114***	0.054***	−0.087***	−0.121***	0.121***	−0.098***	−0.100***	0.005
	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Bins: 10	−0.063***	−0.123***	0.050***	−0.095***	−0.138***	0.128***	−0.098***	−0.115***	0.002
	(0.004)	(0.004)	(0.005)	(0.004)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Num. from Bin: 2	−0.031***	−0.050***	0.028***	−0.039***	−0.057***	0.034***	−0.049***	−0.064***	0.001
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Num. from Bin: 3	−0.043***	−0.088***	0.020***	−0.063***	−0.094***	0.067***	−0.077***	−0.094***	−0.009*
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Num. from Bin: 4	−0.056***	−0.110***	0.010***	−0.085***	−0.124***	0.085***	−0.096***	−0.116***	−0.015***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Num. from Bin: 5	−0.071***	−0.131***	−0.004	−0.104***	−0.152***	0.092***	−0.119***	−0.132***	−0.015***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)
Avg. Bin Values	0.003*	0.011***	−0.001	0.004	0.010***	−0.017***	0.008***	0.010***	0.004
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.003)	(0.003)
Constant	0.926***	0.835***	0.656***	0.821***	0.748***	0.507***	0.672***	0.684***	0.447***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
Observations	93,827	135,552	135,552	95,147	110,412	110,412	97,303	82,285	82,285
Adjusted R^2	0.070	0.173	0.112	0.158	0.190	0.086	0.269	0.221	0.226

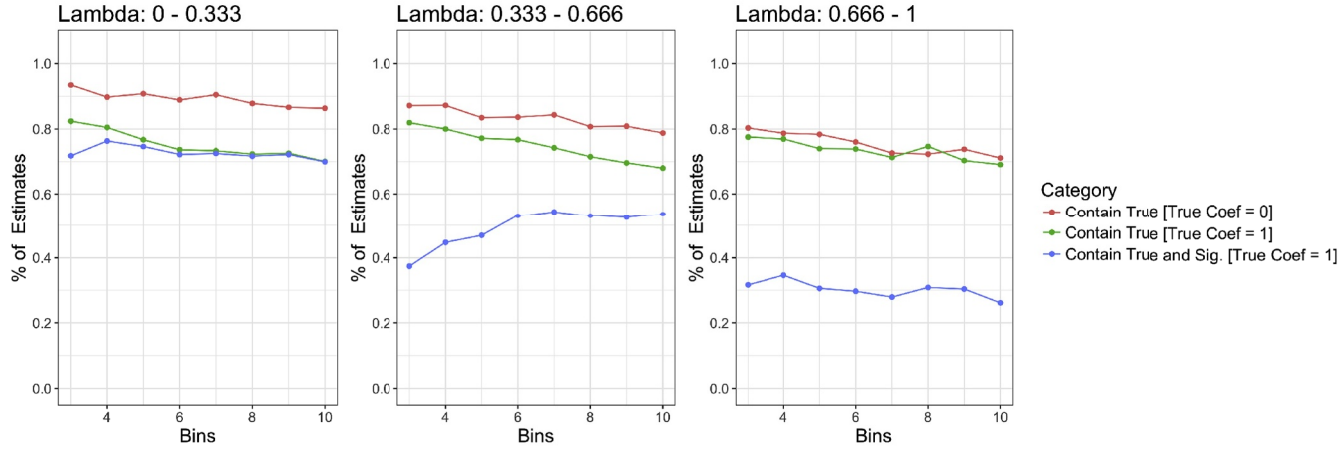
Note: *p < 0.1; **p < 0.05; ***p < 0.01.


Lambda: 0 - 0.333
Lambda: 0.333 - 0.666
Lambda: 0.666 - 1
% of Estimates
Bins
Category
Contain True [True Coef = 0]
Contain True [True Coef = 1]
Contain True and Sig. [True Coef = 1]

(a) 1 Estimate Taken from Each Bin


Lambda: 0 - 0.333
Lambda: 0.333 - 0.666
Lambda: 0.666 - 1
% of Estimates
Bins
Category
Contain True [True Coef = 0]
Contain True [True Coef = 1]
Contain True and Sig. [True Coef = 1]

(b) 2 Estimates Taken from Each Bin


Lambda: 0 - 0.333
Lambda: 0.333 - 0.666
Lambda: 0.666 - 1
% of Estimates
Bins
Category
Contain True [True Coef = 0]
Contain True [True Coef = 1]
Contain True and Sig. [True Coef = 1]

(c) 3 Estimates Taken from Each Bin

Fig. 4. Optimizing GeoSIMEX Parameters.

Fig. 4 illustrates the trade off between the number of bins used and the number of observations drawn from each bin for extrapolations. The y-axis on each chart represents the percentage of all simulations for which a given case was true (contained the correct $\theta = 0$ [red], $\theta = 1$ [green], or $\theta = 1$ with significance [blue] in the 95% confidence interval). The x-axis on each figure is the number of bins used for the GeoSIMEX procedure (i.e., 3C). Three levels of spatial imprecision are shown, with the first column of charts being relatively low imprecision, increasing to high precision in the final column of charts. Each row indicates a different number of observations being selected from each bin for use in back-extrapolation (1, 2, or 3 estimates).

Under all levels of spatial imprecision, increasing the number of bins reduces the ability of GeoSIMEX to capture the true coefficient in the 95% confidence interval. Under low and medium levels of spatial imprecision ($0 \leq \lambda \leq \frac{1}{3}$), increasing the number of bins improves the ability of GeoSIMEX to capture statistical significance. Changes are often largest under low numbers of bins. For example, under low spatial imprecision ($0 \leq \lambda \leq \frac{1}{3}$), moving from using 3 bins to 4 bins (when only one estimate is taken from each bin) results in GeoSIMEX capturing the true coefficient and significance 50% of the time to 65% of the time, while the ability of GeoSIMEX to capture the true coefficient drops from about 91% to 89%. In addition, increasing the number of estimates taken from each bin generally increases the ability of GeoSIMEX to capture both the true coefficient and statistical significance but weakens the ability of GeoSIMEX to capture the true coefficient.

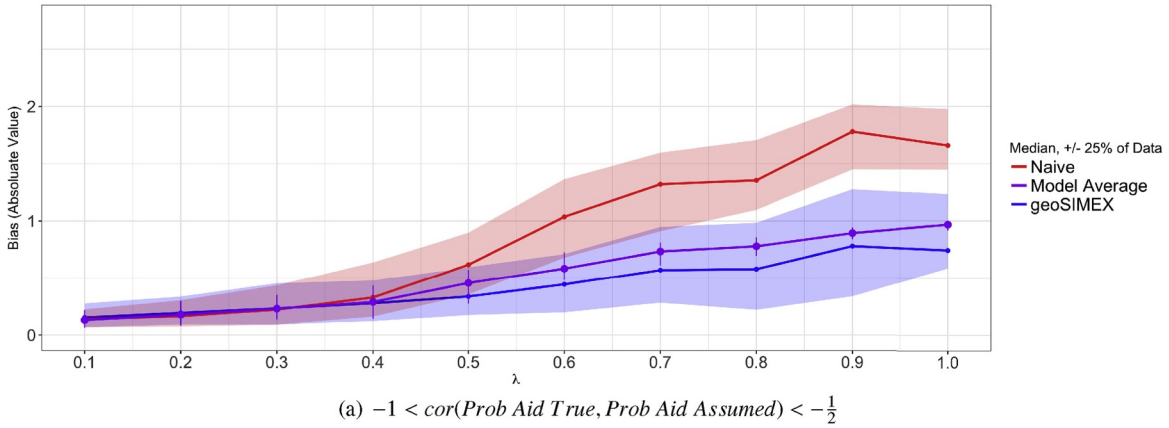
To choose optimal parameters for this analysis (and, to guide future researchers), we calculate the percent of simulation iterations that GeoSIMEX captures the true coefficient and significance across all levels of spatial imprecision, and identify the best parameters for each λ range. Table 5 shows the results of this analysis.

Table 5
Optimal Parameters by Level of Spatial Imprecision.

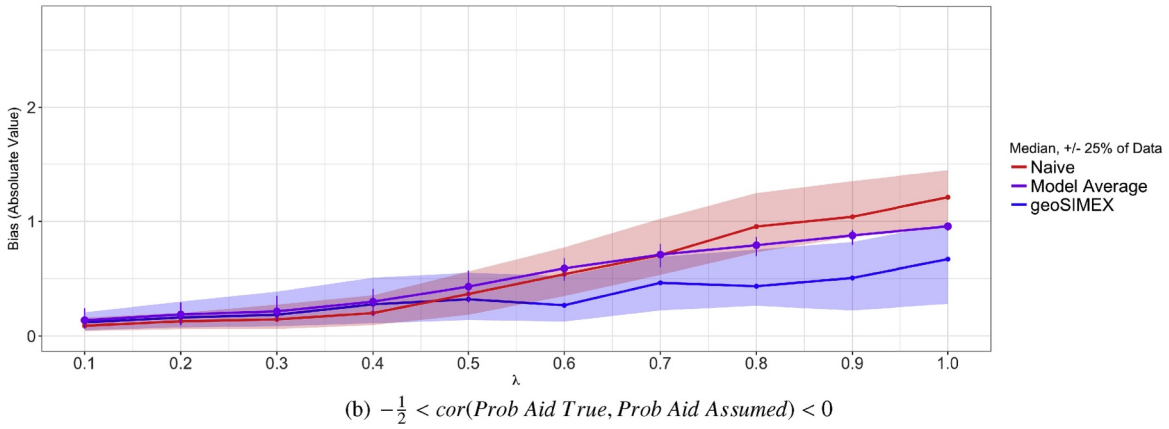
λ	Number of Bins	Number from Bin	GeoSIMEX	
			True	True & Sig.
0–0.1	5	1	0.862	0.837
0.1–0.2	7	1	0.825	0.794
0.2–0.3	8	1	0.812	0.728
0.3–0.4	8	2	0.786	0.695
0.4–0.5	9	5	0.65	0.601
0.5–0.6	10	5	0.635	0.486
0.6–0.7	3	1	0.876	0.403
0.7–0.8	3	1	0.852	0.43
0.8–0.9	4	2	0.748	0.346
0.9–1	4	2	0.649	0.321

A.3. Simulation results: bias and standard errors

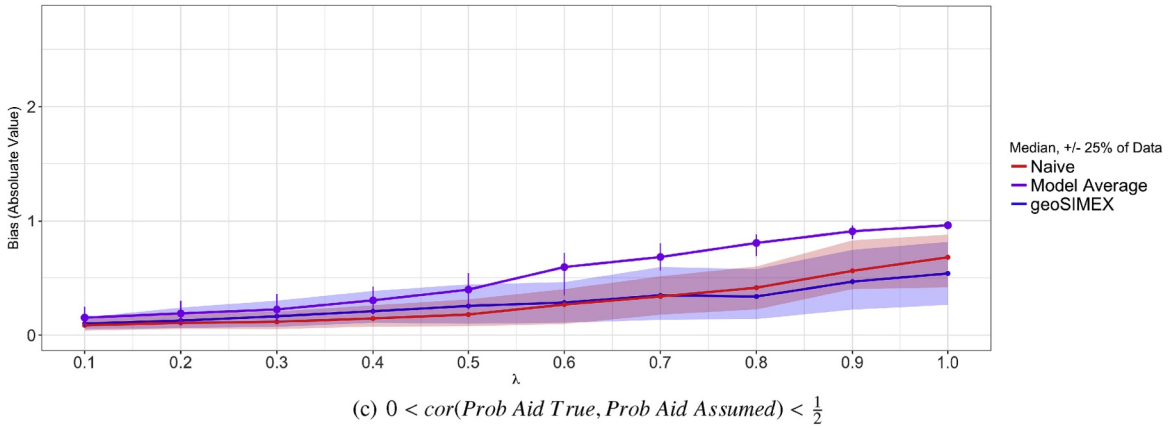
Fig. 5 examines how the bias in θ (y-axis) changes as a function of spatial imprecision (x-axis) under two scenarios. Four scenarios are shown where the accuracy of assumptions about where aid is allocated differ. The first scenario (5a) illustrates how bias can increase as a function of spatial imprecision when very inaccurate assumptions about where aid is allocated are made by the researcher; the fourth scenario (5b) illustrates bias when the researcher makes very accurate assumptions about the spatial allocation of aid in cases of imprecision. Across all scenarios, GeoSIMEX models tend to experience the least amount of bias, especially under conditions of high spatial imprecision.



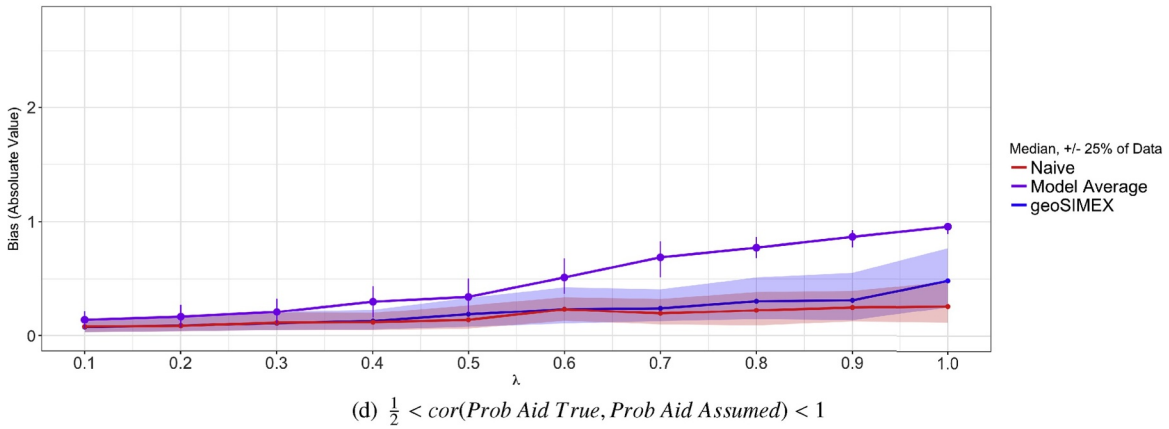
(a) $-1 < cor(Prob\ Aid\ True, Prob\ Aid\ Assumed) < -\frac{1}{2}$



(b) $-\frac{1}{2} < cor(Prob\ Aid\ True, Prob\ Aid\ Assumed) < 0$

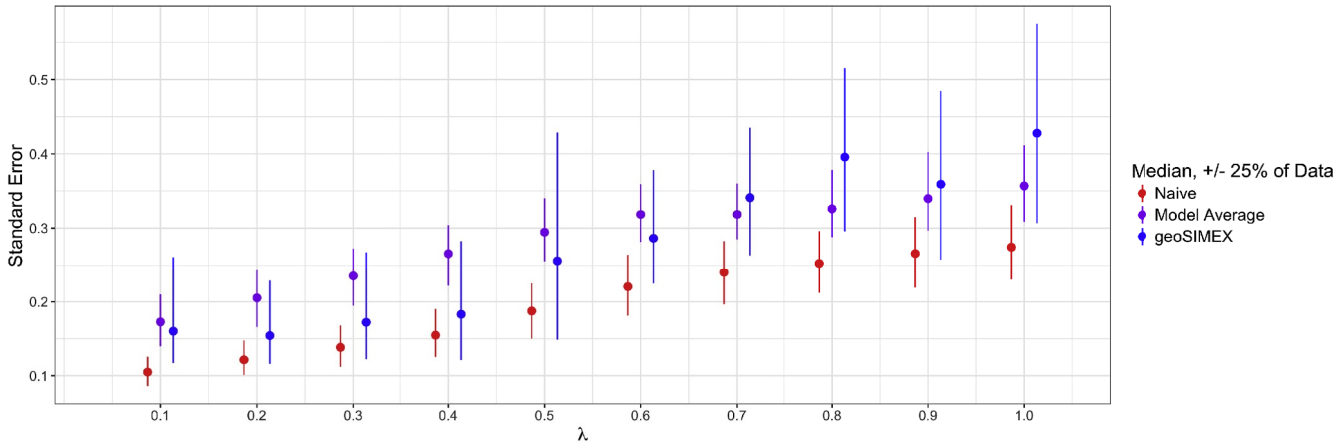


(c) $0 < cor(Prob\ Aid\ True, Prob\ Aid\ Assumed) < \frac{1}{2}$

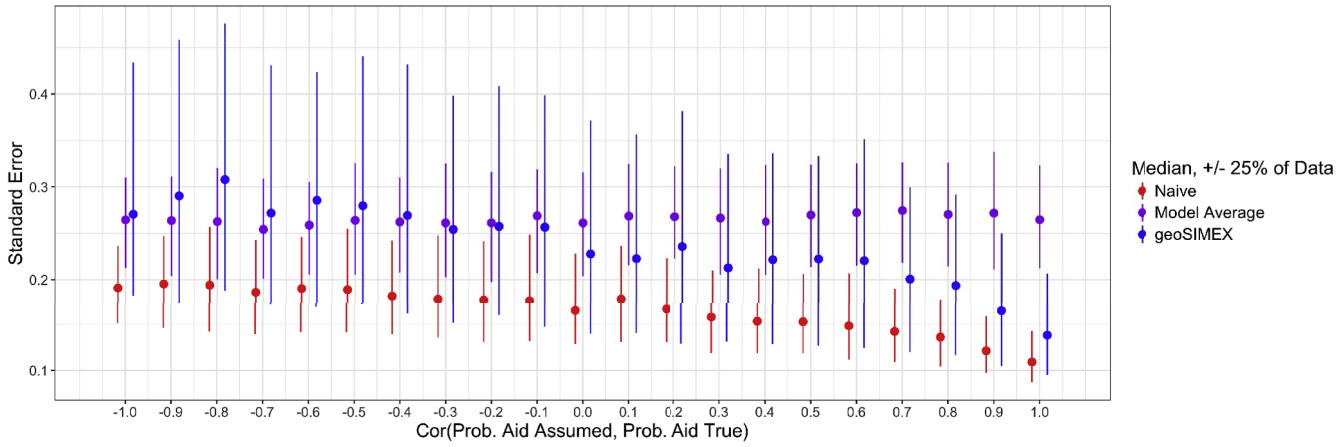


(d) $\frac{1}{2} < cor(Prob\ Aid\ True, Prob\ Aid\ Assumed) < 1$

Fig. 5. Magnitude of Bias. Each line indicates the level of bias of each of three tested models (Naive OLS, Model Averaging, and GeoSIMEX). Regions around each line indicate the 25th percentile (Naive and GeoSIMEX), while 25th percentiles are represented by error bars in the case of Model Averaging for ease of interpretation.



(a) Variation in Standard Errors Across Spatial Imprecision



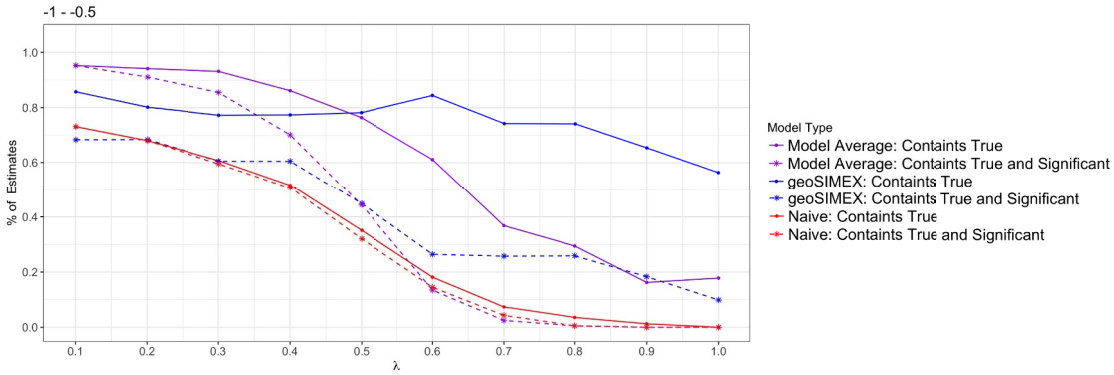
(b) Variation in Standard Errors Across Accuracy of Assumptions

Fig. 6. Variation in geoSIMEX Standard Errors.

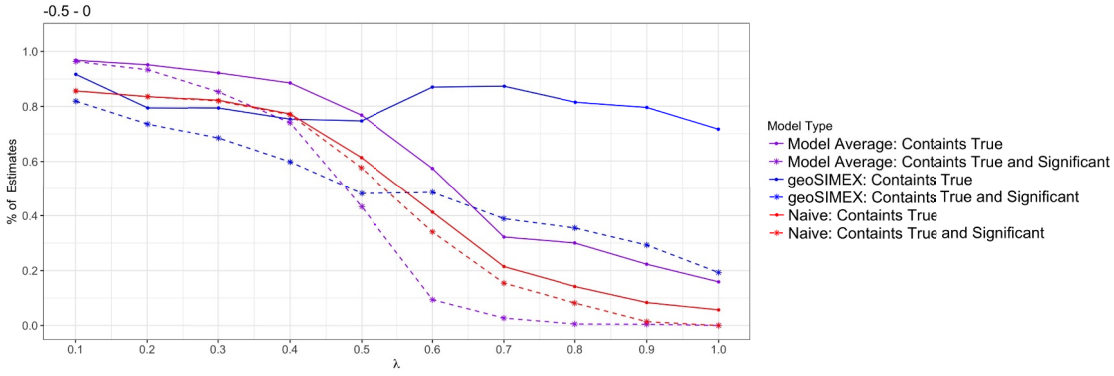
In addition to reducing bias in θ, a key purpose of GeoSIMEX is to incorporate spatial uncertainty into model standard errors. Fig. 6a illustrates how standard errors of aid coefficients increase across all model types—naive, model average and GeoSIMEX. As seen in this figure, GeoSIMEX standard errors grow as the spatial imprecision of the source data increases, following similar trends to Monte Carlo model averaging (though with larger variation, as is expected due to the addition of the back extrapolation step).

Fig. 6b shows the size of model standard errors as quality of aid allocation assumptions change. Larger X-axis values indicate a better assumption being made by the researcher regarding how coarse aid is allocated to finder regions. Standard errors on the naive model get smaller as the accuracy of assumptions improve. GeoSIMEX models also benefit from this reduction in standard error size as the accuracy of assumptions improve (though at a slower rate due to the inclusion of spatial imprecision in standard error estimates).

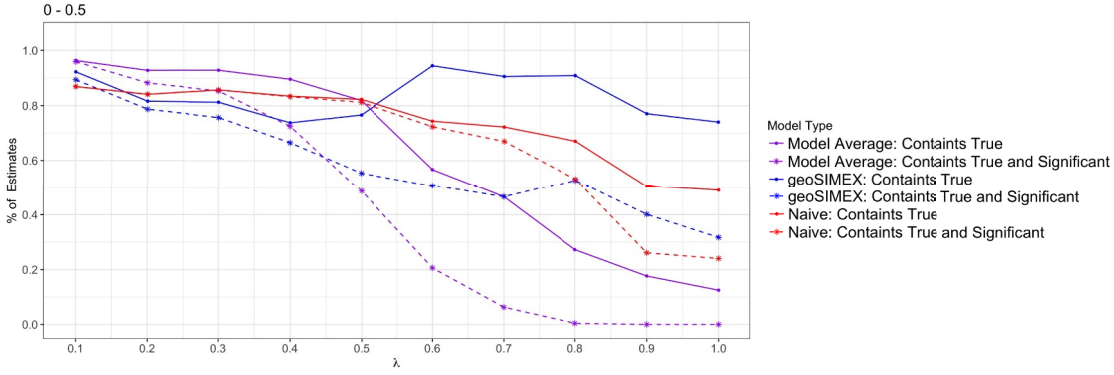
A.4. Performance of GeoSIMEX



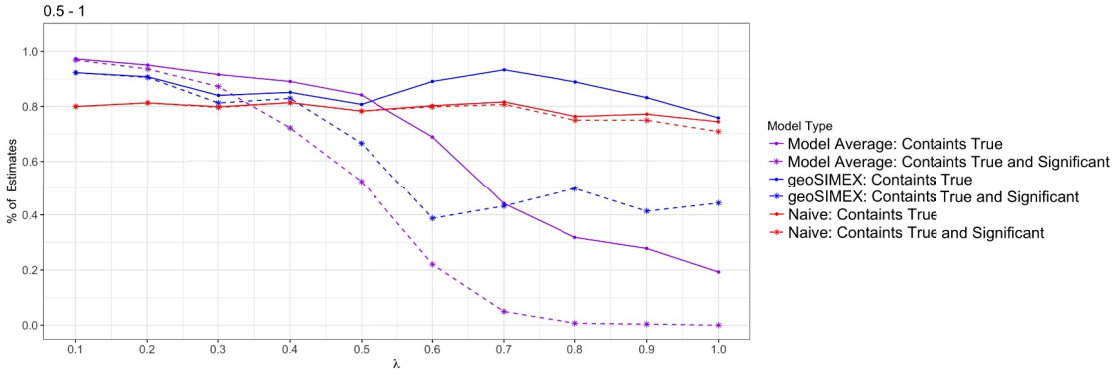
(a) Accuracy of assumptions regarding aid allocation is low (correlated with the truth by -1 - -0.5)



(b) Accuracy of assumptions regarding aid allocation is relatively low (correlated with the truth by -0.5 - 0)



(c) Accuracy of assumptions regarding aid allocation is relatively high (correlated with the truth by 0 - 0.5)



(d) Accuracy of assumptions regarding aid allocation is high (correlated with the truth by 0.5 to 1.0)

Fig. 7. Performance of geoSIMEX, Model Average, and Naive Mode: True Coef (θ) = 1.

Fig. 7 illustrates the performance of GeoSIMEX compared to model averaging and the naive model across spatial imprecision and accuracy of aid allocation assumptions. Four cases are shown - where accuracy of assumptions about aid allocation are low ($r = -1$ - -0.5; 7a), relatively low ($r = -0.5$ - 0; 7b), relatively high (r = 0–0.5; 7c) and high (r = 0.5–1; 7d). The X-axis on each figure represents increasing spatial imprecision, and the Y-axis represents the percentage of simulation iterations for which each condition (represented by each line) was true. Each blue line represents the percentage of times the GeoSIMEX model in particular parameter estimate for θ included the true θ (1) within the 95% confidence interval. The blue dotted line represents when this was true and a significant finding was found. This is repeated for linear models (red lines) and Monte Carlo model averaging (purple lines).

These figures highlight a key advantage of the GeoSIMEX procedure. As spatial imprecision increases, under both cases the ability for traditional model averaging (purple) to identify the true coefficient becomes near or equal to 0. This is due to the fact that, at low levels of precision, Monte Carlo models that do not back-extrapolate will inherently bias their results towards 0 - the attenuation bias referred to earlier in this piece.

Across both levels of assumption accuracy, GeoSIMEX provides similar rates of accuracy to model averaging and linear models for relatively precise (λ ¡ 0.4) datasets. However, as λ values increase, GeoSIMEX retains it's ability to provide accurate parameter estimates at a rate higher than any of the alternative modeling strategies presented here.

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.deveng.2018.11.001.

References

Aerts, Jeroen C.J.H., Goodchild, Michael F., Heuvelink, Gerard B.M., 2003. Accounting for spatial uncertainty in optimization with spatial decision support systems. In: Trans Actions in GIS 7.2, pp. 211–230.

AidData, 2017. Global Chinese Official Finance Dataset. Williamsburg, VA Version 1.0.

Andam, Kwaw S., et al., 2008. “Measuring the effectiveness of protected area networks in reducing deforestation. Proceed. Natl. Acad. Sci. U. S. A. 105 (42), 16089–16094.

Anselin, L., 2002. Under the hood issues in the specification and interpretation of spatial regression models. Agric. Econ. 27 (3), 247–267.

Anselin, L., 2013. Spatial Econometrics: Methods and Models.

Anselin, L., Cho, W.K.T., 2002. Spatial effects and ecological inference. Polit. Anal. 10 (3), 276.

Anselin, Luc, 2003. Spatial externalities. Int. Reg. Sci. Rev. 26 (2), 147–152.

Bare, Matthew, Kauffman, Craig, Miller, Daniel C., 2015. Assessing the impact of international conservation aid on deforestation in sub-Saharan Africa. Environ. Res. Lett. 10 (12), 125010.

BenYishay, A., et al., 2017. Indigenous land rights and deforestation: evidence from the Brazilian Amazon. J. Environ. Econ. Manag.

Bielecka, E., 2005. In: A Dasymetric Population Density Map of Poland, pp. 9–15.

Buchanan, Graeme M., et al., 2016. The impacts of World Bank development projects on sites of high biodiversity importance. AidData Work. Paper Series.

Buntaine, Mark T., Hamilton, Stuart E., Millones, Marco, 2015. Titling community land to prevent deforestation: an evaluation of a best-case program in Morona-Santiago, Ecuador. Global Environ. Change 33, 32–43.

Burnham, K.P., Anderson, D.R., 2002. Information and likelihood theory: a basis for model selection and inference. In: Model Selection and Multimodel Inference: a Practical Information Theoretic Approach, pp. 49–97.

CIESIN, 2000. Gridded Population of the World (GPW). Tech. rep.. SEDAC.

Cook, J.R., Stefanski, L.A., 1994. Simulation-extrapolation estimation in parametric measurement errormodels. J. Am. Stat. Assoc. 89 (428), 1314–1328.

Corrado, Luisa, Fingleton, Bernard, 2012. Where is the economics in spatial econometrics?*. J. Reg. Sci. 52 (2), 210–239.

Dollar, David, Levin, Victoria, 2005. Sowing and reaping: institutional quality and project outcomes in developing countries. SSRN Electron. J.

Dreher, Axel, et al., 2015. Aid on Demand: African Leaders and the Geography of China's Foreign Assistance. Tech. rep. ID 2630152. Social Science Research Network, Rochester, NY.

Drukker, David M., Egger, Peter H., Prucha, Ingmar R., 2013. On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. Econom. Rev. 32 (5–6), 686–733.

Eicher, Cory L., Brewer, Cynthia A., 2001. Dasymetric mapping and areal interpolation: implementation and evaluation. Cartogr. Geogr. Inf. Sci. 28 (2), 125–138.

Foley, Jonathan A., et al., 2005. Global consequences of land use. Sci. New York NY 309 (5734), 570–574.

Gallo, John, Goodchild, Michael, 2012. Mapping uncertainty in conservation assessment as a means toward improved conservation planning and implementation. Soc. Nat. Resour. 25 (1), 22–36.

Grilliches, Z., Hausman, J., 1986. Errors in variables in panel data. Econometrics 93.

Holt, James B., Lo, C.P., Hodler, Thomas W., 2004. Dasymetric estimation of population density and areal interpolation of census data. Cartogr. Geogr. Inf. Sci. 31 (2), 103–121.

Isaksson, Ann-Sofie, Kotsadam, Andreas, 2016. Chinese aid and local corruption. Work. Pap. Econ. 667, 50.

Jones, Stephen G., et al., 2010. Spatial implications associated with using Euclidean distance measurements and geographic centroid imputation in health care research. Health Serv. Res. 45 (1), 316–327.

Kapur, Devesh, Lewis, John P., Webb, Richard C., 1997. The World Bank: its First Half Century: History (English)–The World Bank. Tech. rep.. The World Bank, Washington, DC, pp. 1296.

Kareiva, P., Chang, A., Marvier, M., 2008. Development and conservation goals in World Bank projects. Science 321 (5896), 1638–1639.

LeSage, J., Pace, R.K., 2009. Introduction to Spatial Econometrics.

Li, Yi, Lin, Xihong, 2003. Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. J. Am. Stat. Assoc. 98 (461), 191–203.

McCarthy, J., 2001. IPCC: Schneider and Sarukhan.

McKibbin, Warwick J., Wilcoxen, Peter J., 2002. Climate Change Policy after Kyoto: Blueprint for a Realistic Approach. Brookings Institution Press, pp. 133.

McShane, T.O., et al., 2011. Hard choices:Making trade-offs between biodiversity conservation and human well-being. Biol. Conserv. 144 (3), 966–972.

Nagendra, H., Munroe, D.K., Southworth, J., 2004. From pattern to process: landscape fragmentation and the analysis of land use/land cover change. Agric. Ecosyst. Environ. 101 (2–3), 111–115.

NASA, 2017. LTDR (Land Long Term Data Record) Home. Tech. rep.. NASA, pp. 1.

NOAA, 2017. Earth Observation Group - Defense Meteorological Satellite Program. Tech. rep.. NOAA/NGDC, Boulder, CO.

Ogryczak, Wand, Sliwinski, T., 2009. On effcientWOWA optimization for decision support under risk. Int. J. Approx. Reason. 50 (6), 915–928.

Perez-Heydrich, Carolina, et al., 2013. Guidelines on the Use of DHS GPS Data: DHS Spatial Analysis Reports 8. Tech. rep.. USAID.

Rajan, Raghuram G., Subramanian, Arvind, 2008. Aid and growth: what does the cross-country evidence really show? Rev. Econ. Stat. 90 (4), 643–665.

Rettie, W.J., McLoughlin, P.D., 1999. Overcoming radiotelemetry bias in habitat-selection studies. Canadian J. Zool.-Rev. Canad. Zool. 77 (8), 1175–1184.

Rindfuss, Ronald R., et al., 2004. Developing a science of land change: challenges and methodological issues. Proceed. Natl. Acad. Sci. U. S. A. 101 (39), 13976–13981.

Romero-Lankao, P., et al., 2014. Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II - North America. Tech. rep.. Cambridge University Press, pp. 1439–1498.

Runfola, D., et al., 2016. geoSIMEX: a generalized approach to modeling spatial imprecision. AidData Work. Paper Series 38.

Runfola, D., et al., 2017. A top-down approach to estimating spatially heterogeneous impacts of development aid on vegetative carbon sequestration. Sustainability (Switzerland) 9 (3).

Runfola, Daniel Miller, Napier, Ashley, 2016. Migration, climate, and international aid: examining evidence of satellite, aid, and micro-census data. Migrat. Develop. 5 (2), 275–292.

Shandra, J., Shircliff, E., London, B., 2011. The International Monetary Fund, World Bank, and structural adjustment: a cross-national analysis of forest loss. Soc. Sci. Res. 40 (1), 210–225.

Shandra, John M., 2007. The World polity and deforestation. Int. J. Comp. Sociol. 48 (1), 5–27.

Sreenivas, K., et al., 2014. Spatial assessment of soil organic carbon density through random forests based imputation. J. Indian Soc. Remote Sens. 42 (3), 577–587.

Turner, B.L., Lambin, E.F., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. Proc. Natl. Acad. Sci. 104 (52), 20666.

Turner, W., et al., 2003. Remote sensing for biodiversity science and conservation. Trends Ecol. Evol. 18 (6), 306–314.

Wade, Robert Hunter, 2003. Review: promoting environmental sustainability in development: an evaluation of the World Bank's performance”. Environ. Plann. A 35 (4), 759–760.

Wang, Naisyin, et al., 1998. Bias analysis and SIMEX approach in generalized linear mixed measurement error models. J. Am. Stat. Assoc. 93 (441), 249–261.

Zhao, J., Kemper, P., Runfola, D., 2017a. Quantifying heterogeneous causal treatment effects in World Bank development finance projects. Data Min. Knowl. Discov. (ECML PKDD).

Zhao, J., Kemper, P., Runfola, D., 2017b. In: Simulation Study in Quantifying Heterogeneous Causal Effects.